Effects of Data Errors in Statistical Analysis*

Navya Hooda

February 27, 2024

This paper dissects the impact of instrumental and human errors on statistical analyses.The data used for our purposes follows a Normal Distribution with mean and standard deviation of one and was simulated with certain instrumental errors as a result of human oversight in the cleaning process. More specifically values had decimals shift, and some negative values were turned positive. In this paper we explore the impact of human errors on data analysis through the simulated data and discuss ways to mitigate these errors in statistical analysis.

1 Introduction

Data quality in statistical analysis is important to ensure the accuracy and reliability of the results. The American Statistical Association's Ethical guidelines for statistical practice clearly highlight the need for transparency relating to data limitations, methods, sources of biases when analyzing as a statistical practitioner (American Statistical Association, n.d.). Similarily, we further investigate the errors introduced in the data cleaning process in detail to understand investigate the effects of data errors on statistical analysis using a simulated dataset. The dataset is generated from a normal distribution with a mean of one and a standard deviation of one. However, errors are introduced during data collection and cleaning, including a memory limitation in the instrument that overwrites the final 100 observations with the first 100, and accidental changes to negative values and decimal places. We aim to understand the impact of these errors on statistical analysis and discuss strategies to avoid them.

This paper is structured as follows. In the Data Section, we explore how data was processed. In the Results Section we discuss the findings we discovered after cleaning data and simulation. In the Discussion Section, we address any weaknesses in the data that contribute to our findings, and their impact.

*Code and data are available at: https://github.com/hoodanav/Instrument-Human-Error-Analysis

2 Data

The data was generated to represent a Normal distribution (mean = 1, SD = 1) with 1,000 observations. The last 100 observations were overwritten by the first 100 due to the instrumentation errors we are given. During data cleaning, half of the negative values were made to be positive, and values between 1 and 1.1 had their decimal places shifted by a decimal to the left. The R programming language (R Core Team 2022) was used for both data simulation and cleaning processes. A sample of cleaned data as per errors identified can be seen in Table 1.

Table 1: Snippet of Cleaned Data

data
2.3709584
0.4353018
1.3631284
1.6328626
1.4042683
0.8938755
2.5115220
0.9053410

3 Results

After cleaning the dataset, we find that the mean of the cleaned dataset is approximately 1.015, which is greater than zero. However, this result is biased due to the errors introduced during data collection and cleaning. The duplication of values and modification of decimal places have inflated the mean of the dataset, leading to a biased result.

4 Discussion

The errors introduced in the dataset have a major impact on statistical analysis. The duplication of values and modification of decimal places skew the dataset, leading to biased results. This highlights the importance of data quality in statistical analysis and the need for precise cleaning processes to identify and correct errors. Common strategies to mitigate these errors include documenting data collection and cleaning processes, implementing validation checks through automated processes, peer reviews and conducting sensitivity analyses. These can overall help minimize the effects of data errors on statistical analysis, and reduce the risk of human introduced bias. Ensuring data accuracy is crucial for obtaining reliable insights from data analysis.

References

American Statistical Association. n.d. "Ethical Guidelines for Statistical Practice." https://www.amstat.org/your-career/ethical-guidelines-for-statistical-practice.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.